REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01/05/2007	Progress Report	2/1/07 - 4/30/07

4. TITLE AND SUBTITLE
Next-Generation Image and Sound Processing Strategies: Exploiting the Biological Model

5a. CONTRACT NUMBER
N00014-06-1-0746

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
PI: Mel, Bartlett W.
Co-PI's: Grzywacz, Norberto M., Itti, Laurent, Narayanan, Shri

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Southern California
Dept. of Biomedical Engineering
1042 Downey Way, DRB 140
Los Angeles, CA 90089-1111

8. PERFORMING ORGANIZATION REPORT NUMBER
95-164-2394

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
San Diego Regional Office
140 Sylvester Road
San Diego, CA 92106-3521

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
N00014-06-1-0746

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES
N/A

14. ABSTRACT
Our overarching goal in this project is to extend the technical state-of-the-art in mid-level visual and auditory signal processing using an integrative biologically inspired approach. The work described in this progress report is a continuation of our efforts to (1) imitate biological sensory feature extraction methods, and (2) use those biologically-inspired sensory features to focus attention on the most important information in complex visual and auditory scenes.

15. SUBJECT TERMS
Mid-level visual and auditory processing, biologically inspired, feature extraction, sensory adaptation, attentional focus

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Bartlett W. Mel
none	none	none	none	10	19b. TELEPHONE NUMBER *(Include area code)* 213-740-0334

Progress Report for ONR grant #N00014-06-1-0746

Coverage Period: February 1 through April 30, 2007

Title: Next-Generation Image and Sound Processing Strategies: Exploiting the Biological Model

PI: Bartlett W. Mel, University of Southern California
Co-PI's: Norberto M. Grzywacz, Laurent Itti, Shri Narayanan

Abstract

Our overarching goal in this project is to extend the technical state-of-the-art in mid-level visual and auditory signal processing using an integrative biologically inspired approach. The work described in this progress report is a continuation of our efforts to (1) imitate biological sensory feature extraction methods, and (2) use those biologically-inspired sensory features to focus attention on the most important information in complex visual and auditory scenes. The subsections of this report describe progress in the following specific areas:

1. New experimental measurements of nonlinear center-surround interactions in retinal ganglion cells, comparing responses to artificial stimuli and natural images;

2. Further developments in our biologically-inspired hierarchical feature learning network, including new results on the problem of learning and detecting junction features;

3. Further development of our visual attentional model to incorporate learned top down influences to predict human attentional shifts, plus experimental results;

4. Testing of biologically-inspired feature extraction and normalization operations adapted from our visual attentional model to detect prominent syllables and words in speech;

Recent progress in each of these areas is outlined in the following sections.

1. Comparison of responses of retinal ganglion cells to natural and artificial images

The human visual system beginning at the retina is the best performing image-understanding device known. A major goal of our project is to extract ideas from "wet" experiments and simulation studies of the retina and later visual areas, so that they can be applied to technically difficult image processing problems.

The retina is the first stage of visual processing. Most current models of retinal cells were built to fit responses to artificial stimuli. These well-controlled stimuli include, for example, images of spots, sinusoidal gratings, and white noise. It remains unclear whether models developed using simple artificial stimuli tell the whole story, and in particular, whether key nonlinearities

may be overlooked when identified in these ways. To address this problem, we have set out to compare models of retinal cells characterized with simple stimuli vs. natural images, in the hopes of gaining new insights into nonlinear retinal processing.

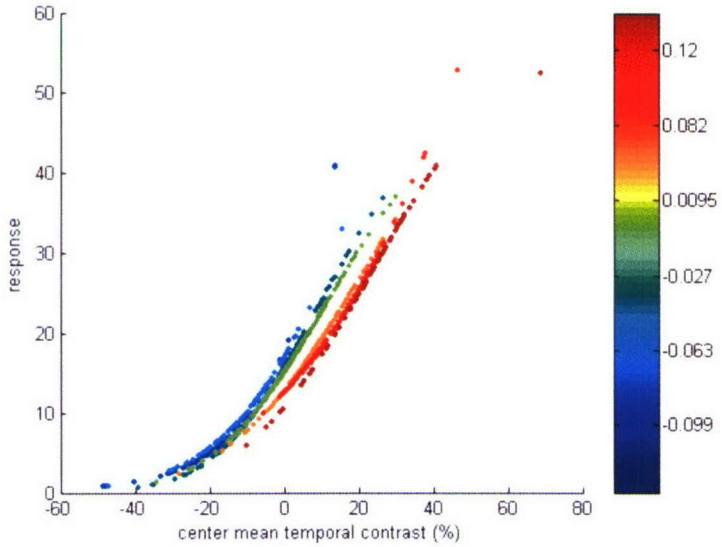

Figure 1. Response-versus-contrast curves for the receptive field center of an ON ganglion cell. Curves are parameterized by the mean contrast of the surround, indicated by the color scale. The slope of the curves is reduced with increasing surround contrast, showing weak division-like inhibition.

We recorded responses from the rabbit retina to either classical artificial stimuli or back-and-forth transitions from homogeneous gray images to randomly drawn natural scenes. Responses to natural images showed properties that could not be predicted from responses to artificial stimuli. We describe three examples: (1) Contrary to classical notions of retinal receptive field organization, surround inhibition was weak for most cells, for all but a few outlier natural images (Figure 1). Furthermore, unlike classical computer vision Mexican hat models of center-surround filters, whose strong surrounds subtractively cancel out the center responses, we found the center-surround interaction was divisive (Figure 1); (2) Cells were either ON or OFF as classically defined by luminance modulation in their receptive-field centers, for both artificial and natural stimuli. However, contrary to expectations from classical descriptions of ON and OFF cells, ON cells responded more strongly to gray-to-natural-image transitions when average luminance was held constant (Figure 2), whereas OFF cells preferred natural-image-to-gray transitions; (3) Some cells were sustained responders with natural images but transient with artificial stimuli.

We are currently trying to incorporate these new findings into improved models of the nonlinear center-surround antagonism in the early stages of visual processing. We will then incorporate

these improved models into our artificial feature extraction systems (discussed in the following sections) to test their efficacy in real world situations.

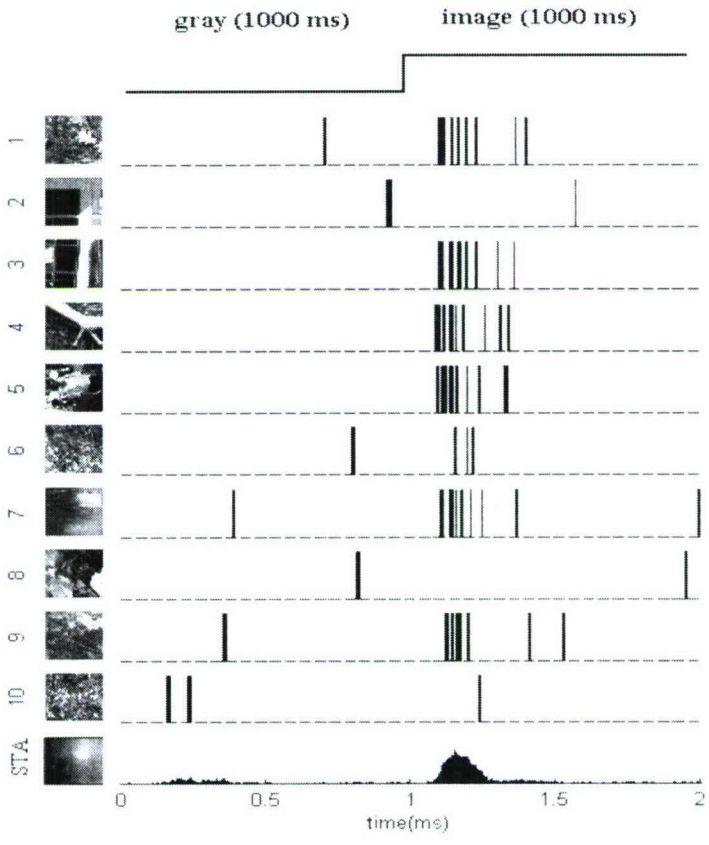

Figure 2. Asymmetric responses of an ON ganglion cell to transitions from image to gray (first 1000 ms) and from gray to image (second 1000 ms). First ten traces correspond to examples of natural-image responses (vertical lines show spike times), while the bottom trace shows the post-stimulus histogram after 1,000 images. This cell responds almost exclusively to gray-to-image transitions.

2. Ongoing developments in a hierarchical feature learning network; applications to the learning of L-junctions

As discussed in previous progress reports, we are developing a hierarchical, nested self-organizing map (SOM) architecture to learn to detect junctions (and eventually objects) in complex visual scenes. Correct classifications of junctions is vital for shape-based object recognition. However, reliably extracting junctions in video images has proven to be very technically challenging. In the last progress report, we showed results from our newly developed end-stop edge detector and corner detectors, the ingredients from which L-junctions are formed. We have now used these two kinds of detectors at the input layer of our nested SOM architecture (Figure 3).

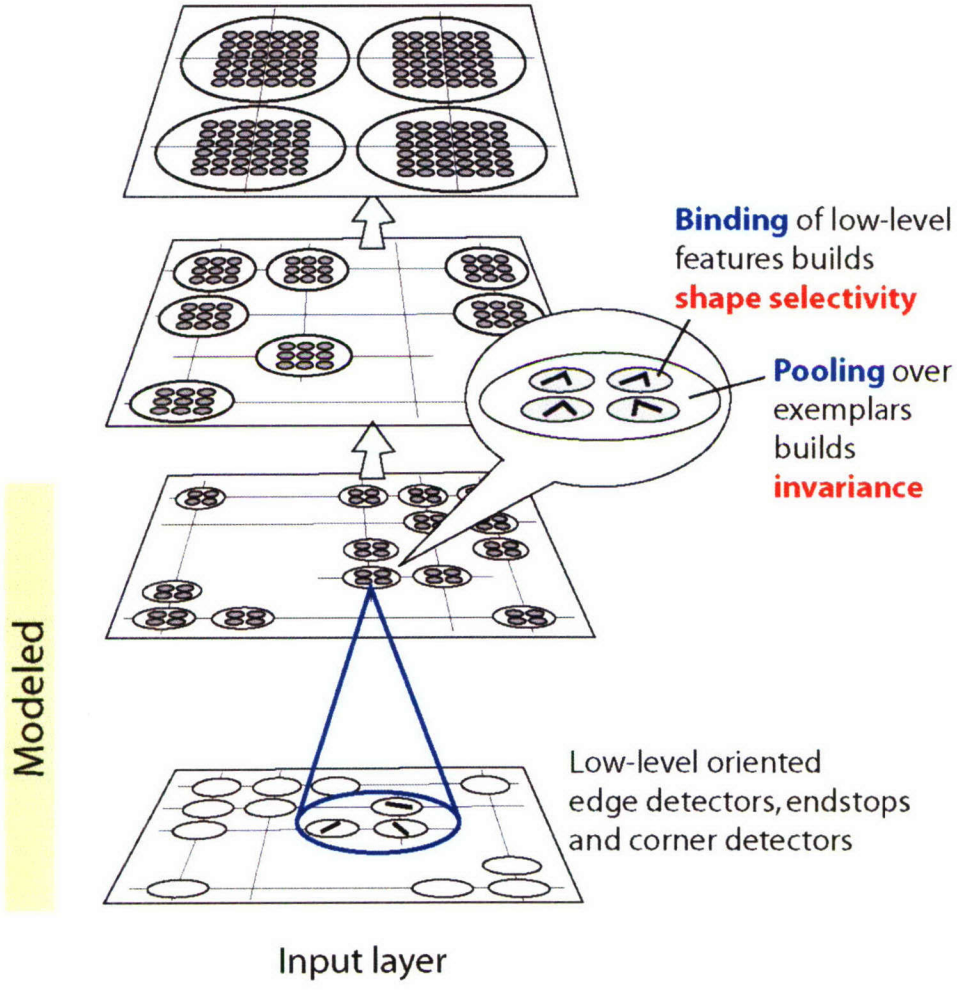

Figure 3. Hierarchical nested self-organizing map (SOM) architecture. The network has so far been modeled through the second layer. Each layer is a self-organizing map trained with a Kohonen-style algorithm. Unlike the standard model, however, the nodes of the SOM are themselves mini-SOMs with several subunits that absorb small spatial variations and deformations. This organization reflects the observation that in the cortical visual processing stream, RFs grow more complex with alternating stages of binding to increase selectivity, and pooling to increase invariance.

We trained the network using 7,000 short movies of corners transforming as if they were observed from a smoothly varying 3-D viewpoint. Beginning in an initial random state, the network was able to discover the underlying structure of the feature space to which it was exposed (Figure 4). Unlike conventional "flat" SOMs, however, the network organized it's learned features in a two-level hierarchy; the coarse level corresponded roughly to a cortical "column" consisting of a complete set of features analyzing one small region of the image. At the fine level, each feature was a "complex" mini-SOM containing 4 subunits, each of which was tuned to a specific L-junction. This pooling within the mini-SOMs led to a modest degree of viewpoint invariance.

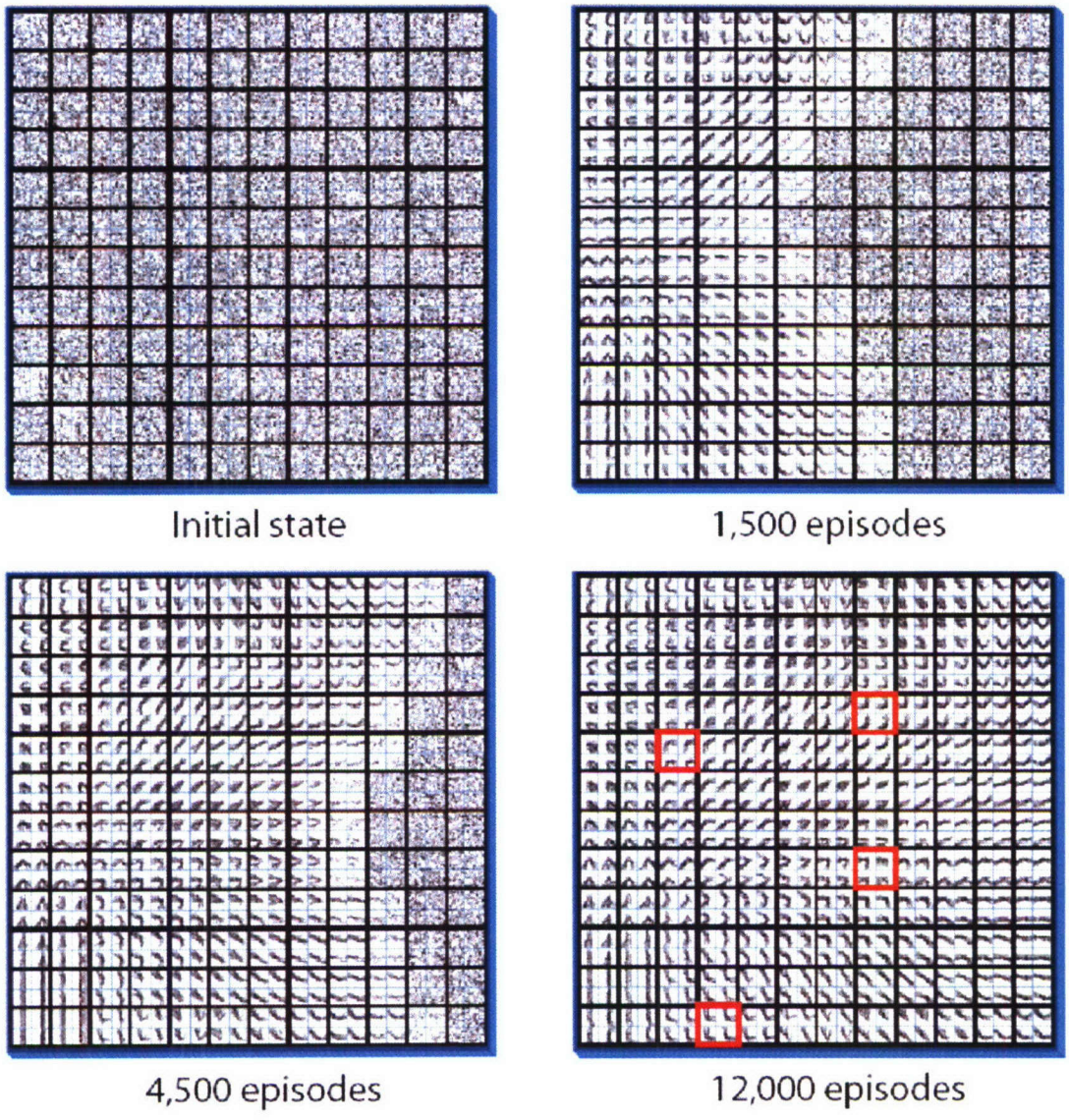

Figure 4. Training of the hierarchical SOM. Network was trained with short movies constructed from 7,000 individual images of L-junctions changing in angle and orientation. Each 2x2 mini-network (outlined with a dark box) is a mini-SOM whose subunits are analogous to simple-cell subunits pooled within a single complex cell. Each mini-SOM, as for the entire network, is trained with a Kohonen-style algorithm. After 12,000 training epochs beginning in a random initial state (upper left), the network converges to a well-formed set of L-junction-sensitive units (lower right). Complex cells outlined in red correspond to the 4 cardinal right-angle corners.

We are currently developing methods for visualizing the responses of the learned complex cell receptive fields when applied to video images, which will allow the properties of the learned L-junction features to be evaluated under varying image conditions. This will allow our learning algorithm to be optimized to maximize the sensitivity and selectivity of the learned features. The

next phases of this project will involve extending the training to multiple contour and junction types, and to implement the second level of the hierarchy capable of learning combinations of junction and contour features.

3. Incorporating top-down influences in attentional selection

A critical function in both machine and biological vision systems is attentional selection of scene regions worthy of further analysis by higher-level processes such as object recognition. (The same is true for auditory attention, as discussed in the next section). In the context of the present project, we have over the past few months brought out the first model of spatial attention that (1) can be applied to arbitrary static and dynamic image sequences with interactive tasks and (2) combines a general computational implementation of both bottom-up (BU) saliency and dynamic top-down (TD) task relevance (Figure 5). The novelty lies in the combination of these elements and in the fully automated nature of the model. The BU component computes a saliency map from 12 low-level multi-scale visual features. (At present the spatial features we use – oriented edges – are simpler than the junction features discussed in the previous section. Given that junctions are highly salient features in natural scenes, however, we plan to incorporate them in future implementations.) The TD component computes a low-level signature of the entire image, and learns to associate different classes of signatures with the different gaze patterns recorded from human subjects performing a task of interest. It is important to note that while we call this component top-down, that does not necessarily imply that it is high-level – in fact, while the learning phase certainly uses high-level information about the scenes, that becomes summarized into the learned associations between scene signatures and expected behavior in the form of TD gaze-position maps.

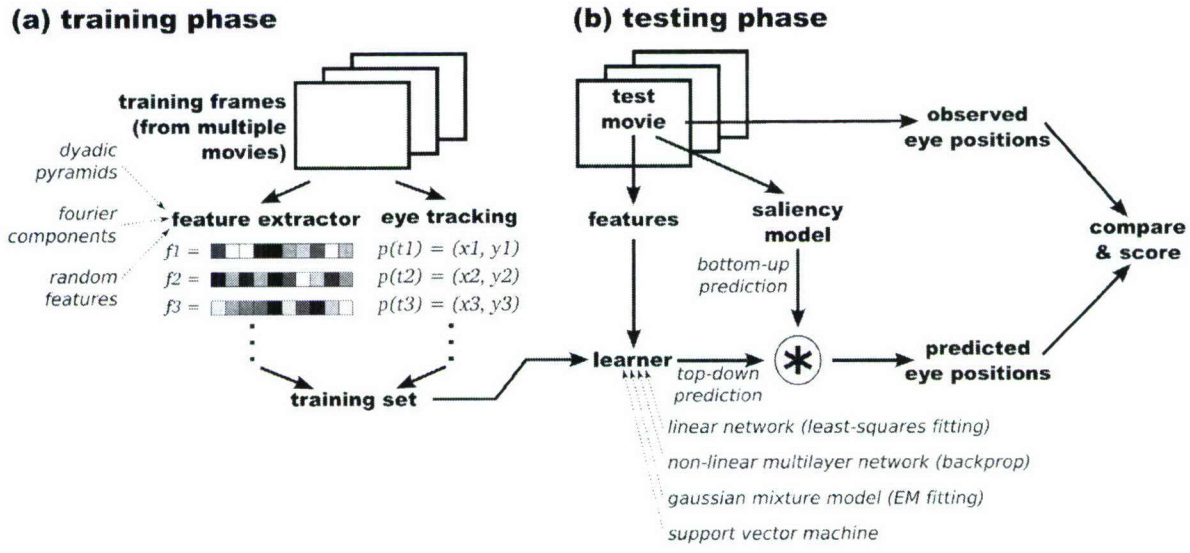

Figure 5. Schematic illustration of our model for learning task-dependent, top-down influences on eye position. First, in (a) the training phase, we compile a training set containing feature vectors and eye positions corresponding to individual frames from several video game clips which were recorded while observers interactively played the games. The feature vectors may be derived from either: the Fourier transform of the image luminance; or dyadic pyramids for luminance, color, and orientation; or as a

control condition, a random distribution. The training set is then passed to a machine learning algorithm to learn a mapping between feature vectors and eye positions. Then, in (b) the testing phase, we use a different video game clip to test the model. Frames from the test clip are passed in parallel to a bottom-up saliency model, as well as to the top-down feature extractor, which generates a feature vector that is used to generate a top-down eye position prediction map. Finally, the bottom-up and top-down prediction maps can be combined via point-wise multiplication, and the individual and combined maps can be compared against the actual observed eye position.

We measured (Peters & Itti, 2007) the ability of this model to predict the eye movements of people playing contemporary video games (Figure 6). We found that the TD model alone predicts where humans look about twice as well as does the BU model alone; in addition, a combined BU*TD model performs significantly better than either individual component. Qualitatively, the combined model predicts some easy-to-describe but hard-to-compute aspects of attentional selection, such as shifting attention leftward when approaching a left turn along a racing track. Thus, our study demonstrates the advantages of integrating bottom-up factors derived from a saliency map and top-down factors learned from image and task contexts in predicting where humans look while performing complex visually-guided behavior. In continuing work we are exploring ways of introducing additional domain knowledge into the top-down component of our attentional system.

- Peters, RJ & Itti, L (2007) Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention, To appear in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

4. Biologically inspired speech and audio processing

As in the case of visual attention, a bottom-up or saliency driven attention allows the brain to detect nonspecific conspicuous targets in cluttered auditory scenes before fully processing and recognizing the targets. We developed a novel biologically inspired auditory saliency map to model such stimulus-driven bottom-up auditory attention. The auditory saliency map builds on the saliency map architecture proposed in Itti & Baldi (2005) for visual attention, and is similar to that used in the visual attention system discussed above.

We tested the auditory saliency map in the context of a "prominent syllable" detection task in speech. The motivation behind choosing prominent syllables is that during speech perception, a particular phoneme or syllable can be perceived to be more salient than the others due to the coarticulation between phonemes, and other factors such as the accent, and physical and emotional state of the talker. This information encoded in the acoustical signal is perceived by the listeners, and we propose to detect these salient syllable locations using the proposed bottom-up auditory attention model.

Then, each feature map is computed by center-surround operation akin to local cortical inhibition. It is implemented in the model by comparing fine and coarse scales. To integrate different features into a single saliency map, a biologically inspired nonlinear local normalization algorithm is used. In our work, the normalization algorithm is adapted from the model proposed for vision to a plausible model for the auditory system. In the next stage, obtained feature maps

are normalized and combined across-scales. At the last stage final maps are summed to output the final auditory saliency map.

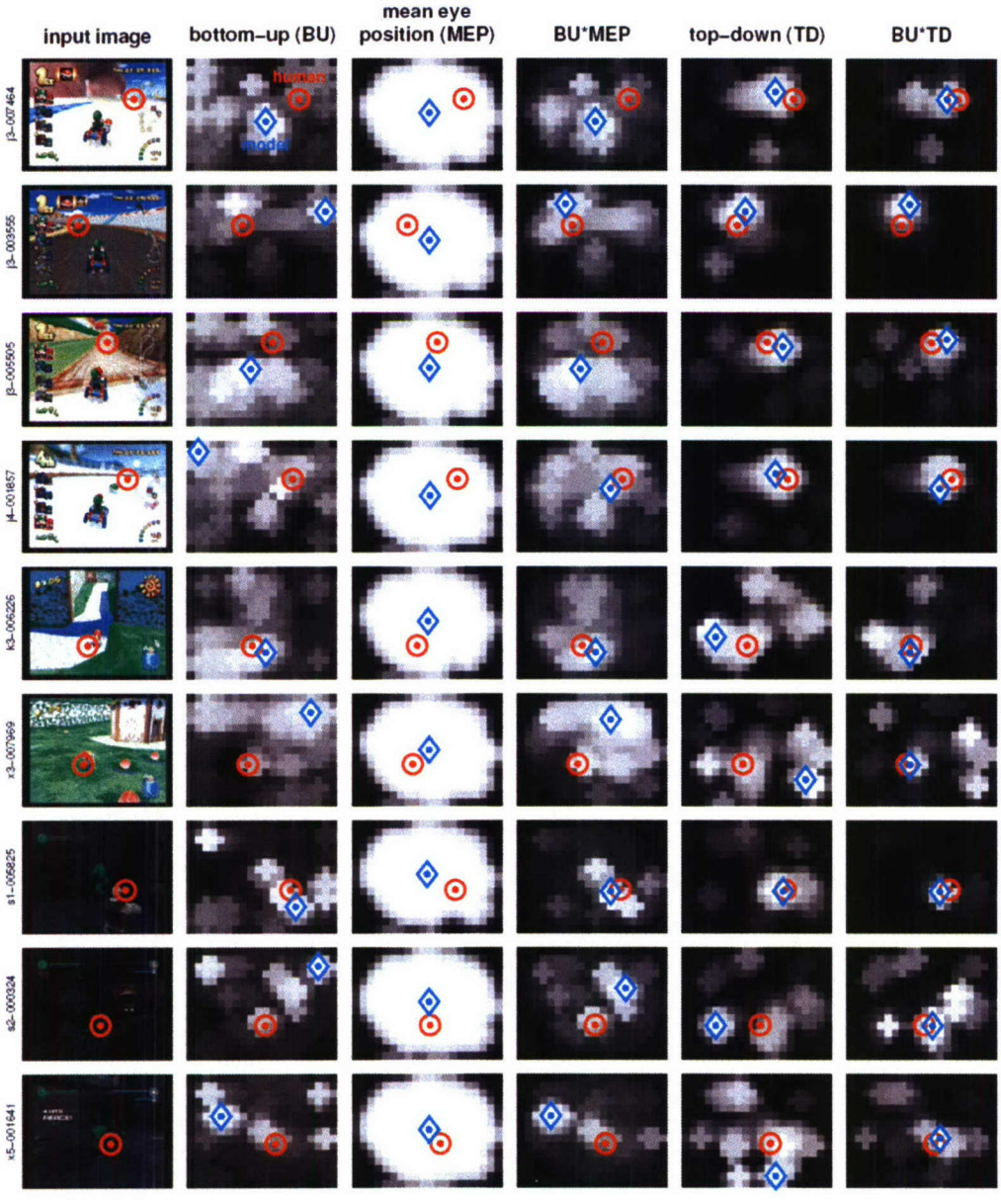

Figure 6. Each row shows a sample video game frame along with the predicted eye position maps generated by several of the computation models that we tested: the purely bottom-up saliency model (BU), the mean eye position prediction (MEP), the point-wise product (BU*MEP), the top-down model based on pyramidal features (TD), and the pointwise product of BU*TD. Each orange circle indicates the observer's actual eye position from that frame; note that this is fixed within each row. Superimposed on the model prediction maps are blue diamonds which indicate the location of each map's peak; a smaller distance between the orange circle and blue diamond suggest a better eye position prediction by the model.

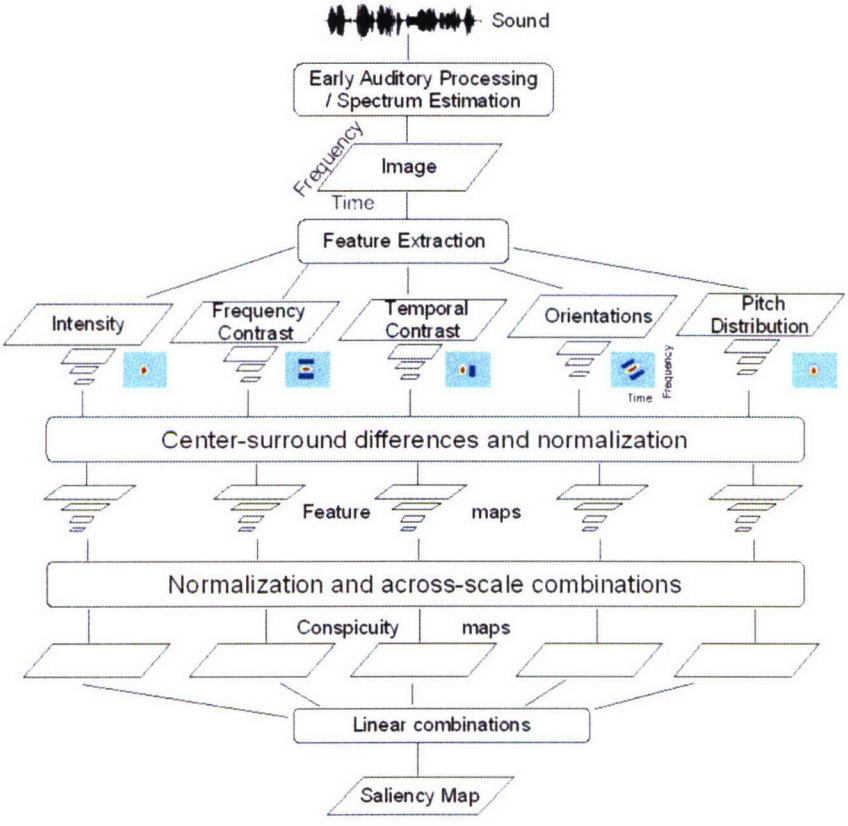

Figure 7. Auditory saliency map structure. Colored images show STRF used for feature extraction

The steps to obtain the auditory saliency map can be summarized as follows: an auditory spectrum of the sound is first computed based on early stages of human auditory system. In the next stage, this spectrum is analyzed by extracting a set of multi-scale features that are similar to the information processing stages in the central auditory system. Intensity, frequency contrast, temporal contrast and orientation features are extracted using spectro-temporal receptive filters (STRF) mimicking the analysis stages in the primary auditory cortex. Pitch is also included in our model, because it is an important property of sound and recent functional imaging studies showed that the neurons of the auditory cortex also respond to pitch.

The maximum of the saliency map defines the most salient location in 2-D auditory spectrum. However, there is neither available saliency ranking for prominent syllables, nor is there information regarding the frequency location that makes the syllable prominent at that time point. We assumed that saliency combines additively across frequency channels, and summing the saliency map across frequency channels for each time point yields a saliency score S(t) for that time point. The local maxima of S(t) which are above a threshold are found, and the syllable at the corresponding time point is marked as prominent.

To test our auditory saliency model, the Boston University Radio News Corpus (BU-RNC) database was used in the experiments. In this database, syllables are stress labeled based on human perception. The syllables annotated with any types of pitch accent were labeled

"prominent", otherwise "non-prominent". Also, we derived word level prominence tags from the syllable level prominence tags. The words that contain one or more prominent syllables are labeled as prominent, non-prominent otherwise. The prominent syllable fraction in the BU-RNC corpus is 34.3% (chance level), and 54.3% (chance level) is the prominent word fraction.

The contribution of each feature to the prominent syllable detection task is examined. The initial letter of the feature names used here to denote the corresponding conspicuity map, i.e. I= Intensity, F=Feature contrast, T=temporal contrast, O=Orientation, P=Pitch. The combination of letters indicates the conspicuity maps that contribute to the saliency map. The best performance is 75.9% accuracy (Acc) with an F-score=0.71, and obtained when the auditory saliency map consisted of I, F, T and O features (IFTO), for the prominent syllable detection task. Even though pitch is an important prosodic cue, the performance obtained with only pitch feature (P) is low (Acc=65.9%), and when it is combined with the rest of the features, it also causes performance degradation (Acc=73.0% with IFTOP). This can be due to two reasons: i) even though the auditory experiments show that human perceive pitch, where/how in the brain it is computed is ambiguous, so the pitch feature may not be modeled correctly in the proposed framework ii) as the findings of a study in the literature, loudness (or intensity here) predicts the syllable prominence, and pitch does not contribute much for syllable prominence task. The word prominence performance is also evaluated similarly. We achieved 78.1% accuracy with an F-score=0.82 for the word prominence task. These results are encouraging given that the average intertranscriber agreement for manual annotators is 80-85% for stress labeling. The results also compare well against the previously reported prosody labeling performance with the BU-RNC.

It can be concluded that the proposed auditory saliency model can successfully detect the prominent syllable and word locations in speech. One advantage of this attention model is that it is language independent, and can detect the prominent syllables in an unsupervised manner. The auditory saliency model proposed here is not only limited to prosody labeling. For example, it can be used in general computational auditory scene analysis (CASA) applications to select conspicuous events rapidly. Similar to the selective attention in humans, after a conspicuous location is selected (focused), it can be analyzed further to recognize the details of the object.

In this work, features are combined with equal weights to create the saliency map. As part of our future work, the weights will be learned in a supervised fashion for different types of auditory tasks, i.e. general audio scene analysis, spoken language processing etc. This can provide insights into what types of cues human brain uses while pre-attending to events in a given task.

- Kalinli, O. & Narayanan, S. "Early Auditory Processing Inspired Features for Robust Automatic Speech Recognition". Submitted to EUROSPEECH '07.

- Kalinli, O. & Narayanan, S. "A Saliency-Based Auditory Attention Model with Applications to Unsupervised Prominent Syllable Detection in Speech". Submitted to INTERSPEECH '07.